On the inexact scaled gradient projection method

Max Lemes

Universidade Federal de Goiás, Brazil

Seminário de Otimização do IME/UFG

Goiânia - Brazil, June 10, 2021

Joint work with: **Orizon P. Ferreira (IME/UFG)**, **Leandro F. Prudente (IME/UFG)**.

Outline

The problem, definitions and preliminaries results

Inexact scaled gradient method

Partial asymptotic convergence

Full asymptotic convergence

Iteration-complexity bound

Numerical experiments

The problem, definitions and preliminaries results

The main problem

We want to present an inexact version of the scaled gradient projection method for constrained convex optimization problem as follows

$$\min\{f(x) : \; x \in C\}, \tag{1}$$

where C is a closed and convex subset of $\mathbb{R}^n$ and $f : \mathbb{R}^n \to \mathbb{R}$ is a continuously differentiable function.

Scaled Gradient Projection Method[1]

Step 0. Choose $\sigma, \tau \in (0,1)$, $0 < \alpha_{\min} \leq \alpha_{\max}$. Let $x^0 \in C$ and set $k = 0$;

Step 1. Choose $\alpha_k \in [\alpha_{\min}, \alpha_{\max}]$ and a positive definite matrix D_k and take $w^k \in C$ as

$$w^k := \mathcal{P}_C^{D_k}(x^k - \alpha_k D_k^{-1} \nabla f(x^k))$$

If $w^k = x^k$, then **stop**; otherwise,

Step 2. Choose τ_k and define the next iterate x^{k+1} as

$$x^{k+1} = x^k + \tau_k(w^k - x^k). \tag{2}$$

and go back to the Step 1.

[1] S. Bonettini, R. Zanella, and L. Zanni. "A scaled gradient projection method for constrained image deblurring". In: *Inverse Problems* 25.1 (2009), pp. 015002, 23.

Scaled Gradient Projection Method

Let D be a $n \times n$ positive definite matrix and $\| \cdot \|_D : \mathbb{R}^n \to \mathbb{R}$ be the norm defined by

$$\|d\|_D := \sqrt{\langle Dd, d \rangle}, \quad \forall d \in \mathbb{R}^n.$$

For a fixed constant $\mu \geq 1$, *denote by* $\mathcal{D}_\mu$ *the set of symmetric positive definite matrices* $n \times n$ *with all eigenvalues contained in the interval* $[\frac{1}{\mu}, \mu]$.

- $\mathcal{D}_\mu$ is compact;
- If $D \in \mathcal{D}_\mu$, it follows that D^{-1} also belongs to $\mathcal{D}_\mu$;
- $\forall D \in \mathcal{D}_\mu$, we obtain

$$\frac{1}{\mu}\|d\|^2 \leq \|d\|_D^2 \leq \mu\|d\|^2, \qquad \forall d \in \mathbb{R}^n.$$

Exact Projection

Definition

The exact projection of the point $v \in \mathbb{R}^n$ onto C with respect to the norm $\|\cdot\|_D$, denoted by $\mathcal{P}_C^D(v)$, is defined by

$$\mathcal{P}_C^D(v) := \arg\min_{z \in C} \|z - v\|_D^2.$$

Lemma

Let $v, w \in \mathbb{R}^n$. Then, $w = \mathcal{P}_C^D(v)$ if and only if $w \in C$ and

$$\langle D(v - w), y - w\rangle \leq 0,$$

for all $y \in C$.

Inexact Projections[2]

Definition
The feasible inexact projection mapping, with respect to the norm $\|\cdot\|_D$, onto C relative to a point $u \in C$ and forcing parameter $\zeta \in (0,1]$, denoted by $\mathcal{P}^D_{C,\zeta}(u,\cdot) : \mathbb{R}^n \rightrightarrows C$, is the set-valued mapping defined as follows

$$\mathcal{P}^D_{C,\zeta}(u,v) := \left\{ w \in C : \ \|w-v\|_D^2 \le \zeta \|\mathcal{P}^D_C(v) - v\|_D^2 + (1-\zeta)\|u-v\|_D^2 \right\}.$$

Each point $w \in \mathcal{P}^D_{C,\zeta}(u,v)$ is called a feasible inexact projection, with respect to the norm $\|\cdot\|_D$, of v onto C relative to u and forcing parameter $\zeta \in (0,1]$.

[2] Ernesto G. Birgin, José Mario Martínez, and Marcos Raydan. "Inexact spectral projected gradient methods on convex sets". In: *IMA J. Numer. Anal.* 23.4 (2003), pp. 539–559.

Inexact Projections[3]

Definition

The feasible inexact projection mapping, with respect to the norm $\|\cdot\|_D$, onto C relative to $u \in C$ and forcing parameter $\gamma \geq 0$, denoted by $\mathcal{R}^D_{C,\gamma}(u,\cdot) : \mathbb{R}^n \rightrightarrows C$, is the set-valued mapping defined as follows

$$\mathcal{R}^D_{C,\gamma}(u,v) := \left\{ w \in C : \ \langle D(v-w), y-w \rangle \leq \gamma \|w-u\|^2_D, \quad \forall\, y \in C \right\}.$$

Each point $w \in \mathcal{R}^D_{C,\gamma}(u,v)$ is called a feasible inexact projection, with respect to the norm $\|\cdot\|_D$, of v onto C relative to u and forcing parameter $\gamma \geq 0$.

[3] Fabiana R. de Oliveira, Orizon P. Ferreira, and Gilson N. Silva. "Newton's method with feasible inexact projections for solving constrained generalized equations". In: *Comput. Optim. Appl.* 72.1 (2019), pp. 159–177.

Inexact Projections

Lemma

Let $v \in \mathbb{R}^n$, $u \in C$, $\gamma \geq 0$ *and* $\zeta \in (0, 1]$. *If* $0 \leq \gamma < 1/2$ *and* $\zeta = 1 - 2\gamma$, *then*

$$\mathcal{R}^D_{C,\gamma}(u, v) \subset \mathcal{P}^D_{C,\zeta}(u, v).$$

Proposition

Let $v \in \mathbb{R}^n$, $u \in C$ *and assume that* C *is a bounded set. Then, for each* $0 < \gamma < 1/2$, *there exist* $0 < \zeta < 1$ *such that*

$$\mathcal{P}^D_{C,\zeta}(u, v) \subseteq \mathcal{R}^D_{C,\gamma}(u, v)$$

Lemma

Let $x \in C$, $\alpha > 0$ *and* $z(\alpha) = x - \alpha D^{-1}\nabla f(x)$. *Take* $w(\alpha) \in \mathcal{P}^D_{C,\zeta}(x, z(\alpha))$ *with* $\zeta \in (0,1]$. *Then, there hold*

(i) $\langle \nabla f(x), w(\alpha) - x\rangle \leq -\frac{1}{2\alpha}\|w(\alpha) - x\|^2_D + \frac{\zeta}{2\alpha}\left[\|\mathcal{P}^D_C(z(\alpha)) - z(\alpha)\|^2_D - \|x - z(\alpha)\|^2_D\right]$;

(ii) *the point* x *is stationary for problem* (1) *if, and only if,* $x \in \mathcal{P}^D_{C,\zeta}(x, z(\alpha))$;

(iii) *if* $x \in C$ *is a nonstationary point for problem* (1), *then* $\left\langle \nabla f(x), w(\alpha) - x \right\rangle < 0$. *Equivalently, if there exists* $\bar{\alpha} > 0$ *such that* $\left\langle \nabla f(x), w(\bar{\alpha}) - x \right\rangle \geq 0$, *then* x *is stationary for problem* (1).

Inexact scaled gradient method

InexProj-SGM employing nonmonotone line search

Step 0. Choose $\sigma, \zeta_{\min} \in (0,1)$, $0 < \alpha_{\min} \le \alpha_{\max}$ and $\mu \ge 1$. Let $x^0 \in C$, $\nu_0 \ge 0$ and set $k \leftarrow 0$.

Step 1. Choose positive real numbers α_k and ζ_k and a positive definite matrix D_k such that

$$\alpha_{\min} \le \alpha_k \le \alpha_{\max}, \qquad 0 < \zeta_{\min} < \zeta_k \le 1, \qquad D_k \in \mathcal{D}_\mu.$$

Compute $w^k \in C$ as any feasible inexact projection with respect to the norm $\|\cdot\|_{D_k}$ of

$$z^k := x^k - \alpha_k D_k^{-1} \nabla f(x^k)$$

onto C relative to x^k with forcing parameter ζ_k, i.e.,

$$w^k \in \mathcal{P}_{C,\zeta_k}^{D_k}(x^k, z^k).$$

If $w^k = x^k$, then **stop** declaring convergence.

InexProj-SGM employing nonmonotone line search

Step 2. Set $\tau_{\rm trial} \leftarrow 1$. If

$$f\Big(x^k + \tau_{\rm trial}(w^k - x^k)\Big) \le f(x^k) + \sigma\tau_{\rm trial}\Big\langle \nabla f(x^k), w^k - x^k\Big\rangle + \nu_k, \tag{3}$$

then $\tau_k \leftarrow \tau_{\rm trial}$, define the next iterate x^{k+1} as

$$x^{k+1} = x^k + \tau_k(w^k - x^k), \tag{4}$$

and go to **Step 3**. Otherwise, choose $\tau_{\rm new} \in [\underline{\omega}\tau_{\rm trial}, \bar{\omega}\tau_{\rm trial}]$, set $\tau_{\rm trial} \leftarrow \tau_{\rm new}$, and repeat test (3).

Step 3. Take $\delta_{k+1} \in [\delta_{\min}, 1]$ and choose $\nu_{k+1} \in \mathbb{R}$ satisfying

$$0 \le \nu_{k+1} \le (1 - \delta_{k+1})\Big[f(x^k) + \nu_k - f(x^{k+1})\Big].$$

Set $k \leftarrow k + 1$ and go to **Step 1**.

Nonmonotone line search

Remarks
There are several ways of choosing ν_k

(i) If $\nu_k = 0$, the line search (4) is the well-known Armijo line search.

(ii) If $f_{\max} = \max\{f(x^{k-j}) \,|\, 0 \leq j \leq \min\{k, M\}\}$ and

$$\nu_k = f_{\max} - f(x^k) \tag{5}$$

the line search (4) is the same defined by Grippo, Lampariello and Lucidi[4].

[4] L. Grippo, F. Lampariello, and S. Lucidi. "A nonmonotone line search technique for Newton's method". In: *SIAM J. Numer. Anal.* 23.4 (1986), pp. 707–716.

Nonmonotone line search

(iii) Let $0 \leq \eta_{min} \leq \eta_{max} < 1$, $c_0 = f(x_0)$ and $q_0 = 1$. Choose $\eta_k \in [\eta_{min}, \eta_{max}]$ and set

$$q_{k+1} = \eta_k q_k + 1, \qquad c_{k+1} = (\eta_k q_k c_k + f(x^{k+1}))/q_{k+1}, \qquad \forall k \in \mathbb{N}.$$

If $\delta_{k+1} = 1/q_{k+1}$ and

$$\nu_k = c_k - f(x^k) \tag{6}$$

the line search (4) is the same defined by Zhang and Hager[5].

[5] H. Zhang and W. W. Hager. "A nonmonotone line search technique and its application to unconstrained optimization". In: *SIAM J. Optim.* 14.4 (2004), pp. 1043–1056.

Partial asymptotic convergence

Partial asymptotic convergence analysis

Lemma

There holds $0 \leq \delta_{k+1}\left[f(x^k) + \nu_k - f(x^{k+1})\right] \leq \left(f(x^k) + \nu_k\right) - \left(f(x^{k+1}) + \nu_{k+1}\right)$*, for all* $k \in \mathbb{N}$*. As consequence the sequence* $\left(f(x^k) + \nu_k\right)_{k\in\mathbb{N}}$ *is non-increasing.*

Theorem

Assume that $\lim_{k\to+\infty} \nu_k = 0$*. Then, Algorithm InexProj-SGM stops in a finite number of iterations at a stationary point of problem* (1)*, or generates an infinite sequence* $(x^k)_{k\in\mathbb{N}}$ *for which every cluster point is stationary for problem* (1)*.*

Partial asymptotic convergence analysis

Proposition

If $\delta_{min} > 0$, *then* $\sum_{k=0}^{+\infty} \nu_k < +\infty$. *Consequently,* $\lim_{k \to +\infty} \nu_k = 0$.[6]

Remark

Armijo line search and nonmonotone line search strategy defined by (6) satisfies a condition $\delta_{min} > 0$. However, for the nonmonotone line search strategy proposed by (5), we can only guarantee that $\delta_{min} \geq 0$. Hence, we need deal with this case separately.

[6] Geovani N. Grapiglia and Ekkehard W. Sachs. "On the worst-case evaluation complexity of non-monotone line search algorithms". In: *Comput. Optim. Appl.* 68.3 (2017), pp. 555–577.

Partial asymptotic convergence analysis

Proposition

Assume that the sequence $(x^k)_{k\in\mathbb{N}}$ *is generated by Algorithm InexProj-SGM with the nonmonotone line search* (5), *i.e.,* $\nu_k = f_{\max} - f(x^k)$ *for all* $k \in \mathbb{N}$*. In addition, assume that the level set* $C_0 := \{x \in C : \ f(x) \leq f(x^0)\}$ *is bounded and* $\nu_0 = 0$*. Then,* $\lim_{k\to+\infty} \nu_k = 0$*.*

Full asymptotic convergence

Full asymptotic convergence analysis

We will prove, under suitable assumptions, the full convergence of the sequence $(x^k)_{k\in\mathbb{N}}$. For this end, we assume that in $\mathbf{Step\,1}$ of Algorithm InexProj-SGM:

A1. For all $k \in \mathbb{N}$, we take $w^k \in \mathcal{R}^{D_k}_{C,\gamma_k}(x^k, z^k)$ with $\gamma_k = (1-\zeta_k)/2$.

A2. For all $k \in \mathbb{N}$, we take $0 \leq \nu_k$ such that $\sum_{k=0}^{+\infty} \nu_k < +\infty$.

Armijo line search and nonmonotone line search strategies defined by (6) satisfies the assumption **A2**.

Full asymptotic convergence analysis

Lemma
For each $x \in C$, there holds

$$\|x^{k+1}-x\|^2_{D_k} \leq \|x^k-x\|^2_{D_k}+2\alpha_k\tau_k\left\langle \nabla f(x^k), x-x^k\right\rangle+\xi\left[f(x^k)-f(x^{k+1})+\nu_k\right], \quad \forall\, k \in \mathbb{N}.$$

where $\xi := \dfrac{2\alpha_{\max}}{\sigma} > 0.$

Corollary
Assume that f is a convex function. If $U := \left\{x \in C : f(x) \leq \inf_{k\in\mathbb{N}} \left(f(x^k)+\nu_k\right)\right\}$ is not empty, then $(x^k)_{k\in\mathbb{N}}$ converges to a stationary point of problem (1).

Full asymptotic convergence analysis

Theorem

If f *is a convex function and* $(x^k)_{k\in\mathbb{N}}$ *has no cluster points, then* $\Omega^* = \varnothing$, $\lim_{k\to\infty} \|x^k\| = +\infty$, *and* $\inf_{k\in\mathbb{N}} f(x^k) = \inf\{f(x) : x \in C\}$.

Full asymptotic convergence analysis

Corollary

If f *is a convex function and* $(x^k)_{k\in\mathbb{N}}$ *has at least one cluster point, then* $(x^k)_{k\in\mathbb{N}}$ *converges to a stationary point of problem* (1).

Full asymptotic convergence analysis

Theorem

Assume that f is a convex function and $\Omega^* \neq \varnothing$. Then, $(x^k)_{k\in\mathbb{N}}$ converge to an optimal solution of problem (1).

Iteration-complexity bound

Interation-complexity bound

Besides assuming that in $\mathbf{Step\,1}$ of Algorithm InexProj-SGM we take $(x^k)_{k\in\mathbb{N}}$ satisfying **A1** and **A2**, we also need the following assumption.

A3. The gradient ∇f of f is Lipschitz continuous with constant $L > 0$.

Lemma
The steepsize τ_k in Algorithm InexProj-SGM satisfies $\tau_k \geq \tau_{\min}$,

where

$$\tau_{\min} := \min\left\{1, \frac{\tau(1-\sigma)}{\alpha_{\max}\mu L}\right\}.$$

Interation-complexity bound

Theorem
For every $N \in \mathbb{N}$, the following inequality holds

$$\min\left\{\|w^k - x^k\| : \; k = 0, 1 \ldots, N-1\right\} \leq \sqrt{\frac{2\alpha_{\max}\mu\left(f(x^0) - f^* + \sum_{k=0}^{\infty}\nu_k\right)}{\sigma\tau_{\min}}}\frac{1}{\sqrt{N}}.$$

Theorem
Let f be a convex function on C. Then, for every $N \in \mathbb{N}$, there holds

$$\min\left\{f(x^k) - f^* : \; k = 0, 1 \ldots, N-1\right\} \leq \frac{\|x^0 - x^*\|_{D_0}^2 + \xi\left[f(x^0) - f^* + \sum_{k=0}^{\infty}\nu_k\right]}{2\alpha_{\min}\tau_{\min}}\frac{1}{N}.$$

Interation-complexity bound

Lemma

Let N_k be the number of function evaluations after $k \geq 1$ iterations of Algorithm InexProj-SGM. Then,

$$N_k \leq 1 + (k+1)\left[\frac{\log(\tau_{\min})}{\log(\tau)} + 1\right].$$

Interation-complexity bound

Theorem

For a given $\epsilon > 0$, the number of function evaluations in Algorithm InexProj-SGM are at most

$$1 + \left(\frac{2\alpha_{\max}\mu\left(f(x^0) - f^* + \sum_{k=0}^{\infty} \nu_k\right)}{\sigma\tau_{\min}} \frac{1}{\epsilon^2} + 1 \right) \left(\frac{\log(\tau_{\min})}{\log(\tau)} + 1 \right),$$

to compute x^k and w^k such that $\|w^k - x^k\| \leq \epsilon$.

Interation-complexity bound

Theorem

Let f be a convex function on C. For a given $\epsilon > 0$, the number of function evaluations in Algorithm InexProj-SGM are at most

$$1 + \left(\frac{\|x^0 - x^*\|_{D_0}^2 + \xi \left(f(x^0) - f^* + \sum_{k=0}^{\infty} \nu_k \right)}{2\alpha_{\min}\tau_{\min}} \frac{1}{\epsilon} + 1 \right) \left(\frac{\log(\tau_{\min})}{\log(\tau)} + 1 \right),$$

to compute x^k such that $f(x^k) - f^* \leq \epsilon$.

Numerical experiments

Numerical experiments

Given A and B two $m \times n$ matrices, with $m \geq n$, and $c \in \mathbb{R}$, we consider the matrix function $f : \mathbb{R}^{n\times n} \to \mathbb{R}$ given by:

$$f(X) := \frac{1}{2}\|AX - B\|_F^2 + \sum_{i=1}^{n-1}\left[c\left(X_{i+1,i+1} - X_{i,i}^2\right)^2 + (1 - X_{i,i})^2\right],$$

which combines a least squares term with a Rosenbrock-type function. $X_{i,j}$ stands for the ij-element of the matrix X and $\|\cdot\|_F$ denotes the Frobenius matrix norm, i.e., $\|A\|_F := \sqrt{\langle A, A\rangle}$ where the inner product is given by $\langle A, B\rangle = \mathrm{tr}(A^T B)$.

Problem I[7]**:**

$$\begin{array}{ll} \min & f(X) \\ \text{s.t.} & X \in SDD^+, \\ & L \leq X \leq U, \end{array}$$

where SDD^+ is the cone of symmetric and diagonally dominant real matrices with positive diagonal, i.e.,

$$SDD^+ := \{X \in \mathbb{R}^{n\times n} \mid X = X^T,\ X_{i,i} \geq \sum_{j\neq i} |X_{i,j}|\ \forall i\},$$

L and U are given $n \times n$ matrices, and $L \leq X \leq U$ means that $L_{i,j} \leq X_{i,j} \leq U_{i,j}$ for all i, j.

[7] Ernesto G. Birgin, José Mario Martínez, and Marcos Raydan. "Inexact spectral projected gradient methods on convex sets". In: *IMA J. Numer. Anal.* 23.4 (2003), pp. 539–559.

Problem II[8,9]**:**

$$\begin{array}{ll} \min & f(X) \\ \text{s.t.} & X \in \mathbb{S}^n_+, \\ & \operatorname{tr}(X) = 1, \end{array}$$

where $\mathbb{S}^n_+$ is the cone of symmetric and positive semidefinite real matrices. The feasible set of Problem II was known as *spectrahedron* and appears in several interesting applications.

[8] Zeyuan Allen-Zhu et al. "Linear convergence of a Frank-Wolfe type algorithm over trace-norm balls". In: *Advances in Neural Information Processing Systems*. 2017, pp. 6191–6200.
[9] D.S. Gonçalves, M.A. Gomes-Ruggiero, and C. Lavor. "A projected gradient method for optimization over density matrices". In: *Optimization Methods and Software* 31.2 (2016), pp. 328–341.

Numerical experiments

We are interested in the spectral gradient version of the SPG method, so we set $D_k := I$ for all k, $\alpha_0 := \min(\alpha_{\max}, \max(\alpha_{\min}, 1/\|\nabla f(x^0)\|))$ and, for $k > 0$,

$$\alpha_k := \begin{cases} \min(\alpha_{\max}, \max(\alpha_{\min}, \langle s^k, s^k\rangle / \langle s^k, y^k\rangle)), & \text{if } \langle s^k, y^k\rangle > 0 \\ \alpha_{\max}, & \text{otherwise,} \end{cases}$$

where $s^k := X^k - X^{k-1}$, $y^k := \nabla f(X^k) - \nabla f(X^{k-1})$, $\alpha_{\min} = 10^{-10}$, and $\alpha_{\max} = 10^{10}$.

Concerning the stopping criterion, all runs were stopped at an iterate X^k declaring convergence if

$$\max_{i,j}(|X^k_{i,j} - W^k_{i,j}|) \leq 10^{-6},$$

where $W^k \in \mathcal{P}^{D_k}_{C,\zeta_k}(x^k, z^k)$.

Influence of the inexact projection

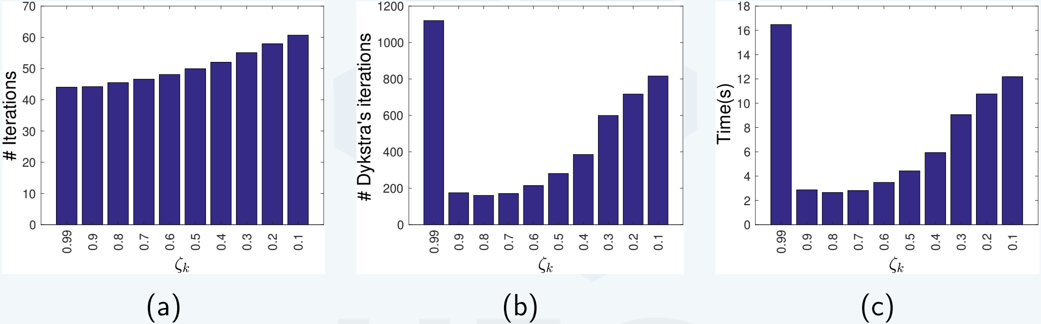
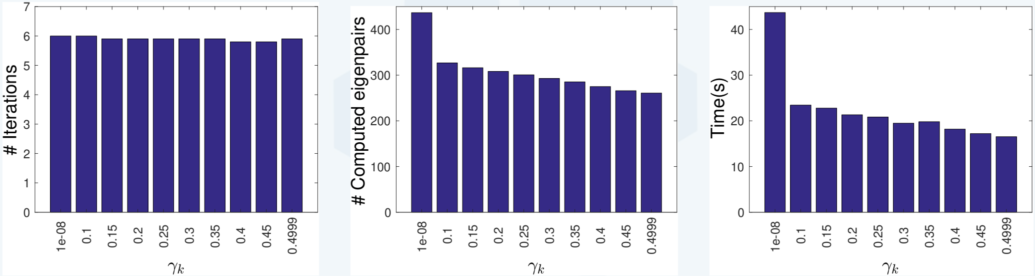


Figure 1: Results for 10 instances of Problem I using $n = 100$, $m = 200$, and $c = 10$. Average number of: (a) iterations; (b) Dykstra's iterations; (c) CPU time in seconds needed to reach the solution for different choices of ζ_k.

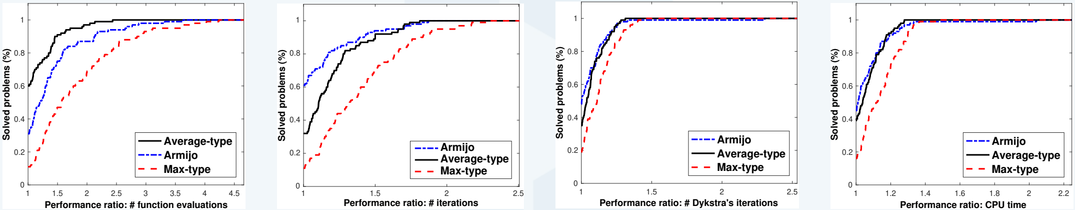
Influence of the inexact projection



(a) (b) (c)

Figure 2: Results for 10 instances of Problem II using $n = 800$, $m = 1000$, and $c = 100$. Average number of: (a) iterations; (b) computed eigenpairs; (c) CPU time in seconds needed to reach the solution for different choices of γ_k.

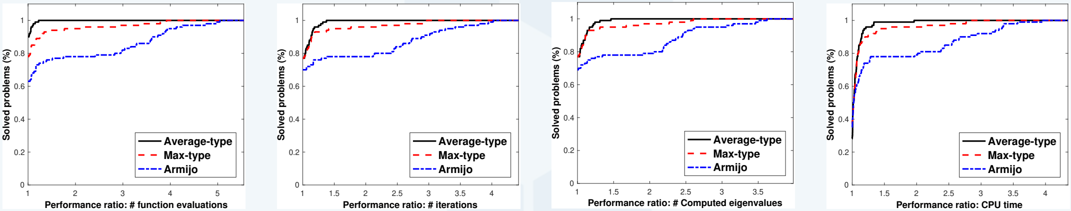
Influence of the line search scheme



(a) (b) (c) (d)

Figure 3: Performance profiles for Problem I considering the SPG method with the Armijo, the Average-type, and the Max-type line searches strategies using as performance measurement: (a) number of function evaluations; (b) number of (outer) iterations; (c) number of Dykstra's iterations; (d) CPU time.

Influence of the line search scheme



(a) (b) (c) (d)

Figure 4: Performance profiles for Problem II considering the SPG method with the Armijo, the Average-type, and the Max-type line searches strategies using as performance measurement: (a) number of function evaluations; (b) number of (outer) iterations; (c) number of computed eigenpairs; (d) CPU time.

Thank you!